Developing Guidelines and Ensuring Consistency for Chinese Text Annotation

Fei Xia[*], Martha Palmer[*], Nianwen Xue[§], Mary Ellen Okurowski[¶], John Kovarik[¶],
Fu-Dong Chiou[†], Shizhe Huang[‡], Tony Kroch[†], Mitch Marcus[*]

[*] Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
{fxia,mpalmer,mitch}@linc.cis.upenn.edu

[†] Linguistics Department
University of Pennsylvania
Philadelphia, PA 19104, USA
{chioufd,kroch}@linc.cis.upenn.edu

[§] Linguistics Department
University of Delaware
Newark, DE 19716, USA
xueniwen@UDel.Edu

[¶] US Department of Defense
Ft. Meade, MD 20755, USA
meokuro@super.org, kovariks@worldnet.att.net

[‡] East Asian Studies Program
Haverford College
Haverford, PA 19041, USA
shuang@haverford.edu

Abstract

With growing interest in Chinese Language Processing, numerous NLP tools (e.g. word segmenters, part-of-speech taggers, and parsers) for Chinese have been developed all over the world. However, since no large-scale bracketed corpora are available to the public, these tools are trained on the corpora with different segmentation criteria, part-of-speech tagsets and bracketing guidelines, and therefore, comparisons are difficult. As a first step towards addressing this issue, we have been preparing a 100-thousand-word bracketed corpus since late 1998 and plan to release it to the public summer 2000. In this paper, we will address several challenges in building the corpus, namely, creating annotation guidelines, ensuring annotation accuracy and maintaining a high level of community involvement.

1. Introduction

With growing interest in Chinese Language Processing, numerous NLP tools (e.g. word segmenters, part-of-speech taggers, and parsers) for Chinese have been developed all over the world. However, since there are no standard reference treebank corpora of hand-parsed sentences for Chinese, it is difficult to compare results and gauge progress in the field. As a first step towards addressing this issue, we have been preparing a 100-thousand-word bracketed corpus since late 1998 and plan to release it to the public in the summer of 2000.

In this paper, we will describe several challenges in creating a Chinese treebank and our response to them.

- guideline preparation: preparing good guidelines for word segmentation, part-of-speech tagging and bracketing.
- quality control: ensuring inter-annotator consistency and adherence to the guidelines.
- community involvement: maintaining a high level of community involvement in the project so that the final guidelines and annotated corpora are as widely useful as possible.

We begin with a discussion of community involvement because of its importance at all stages to the creation of a shareable resource. We then overview the project (Section 3), outline our methodology for guideline preparation (Section 4), and detail the word segmentation and POS tagging phase (Sections 5) and the bracketing phase (Sections 6-7).

2. Project Inception

Our first step in assessing community interest in a standard reference corpus for Chinese was a three-day workshop on issues in Chinese language processing and translation which was held at Penn. The aim of this workshop was to bring together influential researchers from Taiwan, Singapore, Hong Kong, China and the United States in a move towards consensus building with respect to word segmentation, part-of-speech (POS) tagging, syntactic bracketing and other areas. The American groups included the Institute for Research in Cognitive Science and the Linguistics Data Consortium (which distributes the English Treebank) at the University of Pennsylvania, the University of Maryland, Queens College, the University of Kansas, the University of Delaware, Johns Hopkins University, Systran, BBN, ATT, Xerox, West, Unisys and the US Department of Defense. We also invited representatives of ROCLING in Taiwan and Hong Kong Science and Technology University. The workshop included presentations of guidelines being used in mainland China, Taiwan, and Hong Kong, as well as segmenters, part-of-speech taggers and parsers. There were also several working groups that discussed specific issues in segmentation, POS tagging and the syntactic annotation of newswire text.[1]

There was general consensus at this workshop that a large-scale effort to create a Chinese Treebank would be well received, and that linguistics expertise was a necessary prerequisite to successful completion of such a project. The workshop made considerable progress in defining criteria for segmentation guidelines as well as addressing the issues of part-of-speech tagging and syntactic bracketing. The Penn Chinese Treebank project began shortly after the workshop was held.[2]

3. Project Overview

Our goal is the creation of a 100,000 word corpus of Chinese with syntactic bracketing. The corpus includes 329 articles from the Xinhua newswire. The majority of these documents (a total of 291) focus on economic developments during April, August, and September of 1994 (121), February through April of 1996 (40), March, April, and December of 1997 (76), and January of 1998 (54), while in an attempt to broaden coverage an additional 38 documents across the same period were annotated which describe general political and cultural topics. Our relatively small corpus contains 173,981 *hanzi* (or 100,996 words after word segmentation). The corpus has 3,289 sentences,[3] averaging 47 *hanzi* (or 27 words after segmentation) per sentence.

The project has two phases: the first phase is word segmentation and part-of-speech (POS) tagging and the second phase is syntactic bracketing. Table 1 shows an example and what it looks like after each phase.[4] At each phase, all the data are annotated at least twice with a second annotator correcting the output of the first annotator. During the process, we have held several meetings to get feedback from the Chinese NLP community and have revised our guidelines accordingly. Figures 1 and 2 summarize the milestones of the project.

(a) Raw data:

他还提出一系列具体措施和政策要点。

(b) After Phase I:

他/PN 还/AD 提出/VV 一/CD 系列/M 具体/JJ
He also propose one series concrete
措施/NN 和/CC 政策/NN 要点/NN 。/PU
measure and policy essential .
(He also proposed a series of concrete
measures and essentials on policy.)

(c) After Phrase II:

```
(IP (NP-SBJ (PN 他))
    (VP (ADVP (AD 还))
        (VP (VV 提出)
            (NP-OBJ (QP (CD 一)
                        (CLP (M 系列)))
                    (NP (NP (ADJP (JJ 具体))
                            (NP (NN 措施)))
                        (CC 和)
                        (NP (NN 政策)
                            (NN 要点))))))
    (PU 。))
```

Table 1: A sample sentence from the corpus

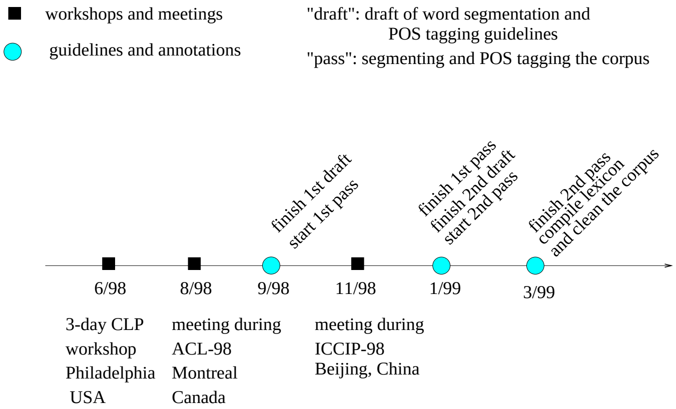
■ workshops and meetings
● guidelines and annotations
"draft": draft of word segmentation and POS tagging guidelines
"pass": segmenting and POS tagging the corpus
finish 1st draft
start 1st pass
finish 1st pass
finish 2nd draft
start 2nd pass
finish 2nd pass
compile lexicon
and clean the corpus
6/98
8/98
9/98
11/98
1/99
3/99
3-day CLP workshop Philadelphia USA
meeting during ACL-98 Montreal Canada
meeting during ICCIP-98 Beijing, China

Figure 1: The first phase: segmentation and POS tagging

[1] The workshop was a small-scale version of the 1992 UPenn meeting of the 35 designers of English-language grammatical analyzers from the United States, Great Britain, and the European continent (Black et al., 1993).

[2] Our Penn Chinese Treebank website, http://www.ldc.upenn.edu/ctb , includes segmentation, POS tagging and bracketing guidelines, as well as sample files, information on our first workshop and much more.

[3] A *sentence* is anything that ends with a period, a exclamation mark or a question mark, therefore, it does not include the headline at the beginning of each article.

[4] The gloss in Table 1(b) is not part of the annotation. It is included here for non-Chinese speakers.

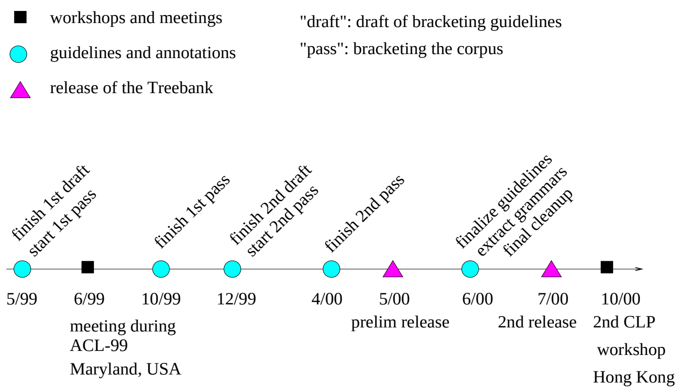

Figure 2: The second phase: bracketing and data release

Our team includes two linguists, three computational linguists, two annotators and several external consultants.

4. Methodology for Guideline Preparation

To create a treebank for Chinese we need to create three sets of guidelines — segmentation, part-speech tagging and bracketing guidelines. Making these guidelines is especially challenging because:

- Unlike Western writing systems, Chinese writing does not have a natural delimiter between words, and the notion of *word* is very hard to define.
- Chinese has very little, if any, inflectional morphology. Words are not inflected with number, gender, case, or tense. For example, a word such as 毁灭 in Chinese corresponds to *destroy/destroys/destroyed/destruction* in English. This fuels the discussion on whether the POS tags should be based on meaning or on syntactic distribution in Chinese NLP communities. If only the meaning is used, 毁灭 should be a verb all the time. If syntactic distribution is used, the word is a verb or a noun depending on the context.
- There are many open questions in Chinese syntax. To further complicate the situation, Chinese, like any other language, is under constant change. With its long history, a seemingly homogeneous phenomenon in Chinese (such as long and short *bei*-construction) may be, in fact, a set of historically related but syntactically independent constructions (Feng, 1998).
- Chinese is widely spoken in areas as diverse as China, Hong Kong, Taiwan, and Singapore. There is a growing body of research in Chinese natural language processing, but little consensus on linguistic standards along the lines of the EAGLES initiative in Europe.[5]

To tackle these issues, we adopted the following approach:

- In addition to studying the literature on Chinese morphology and syntax, we collaborate closely with our linguistics experts to work out plausible analyses for syntactic constructions.
- When there are no clear winners among several alternatives, we choose one, and annotate the corpus in a way that our annotation can be easily converted to accommodate other alternatives when needed.
- We study other groups' guidelines, such as the Segmentation Standard in China (Liu et al., 1993) and the one in Taiwan (Chinese Knowledge Information Processing Group, 1996), and accommodate them in our guidelines if possible.
- We organize regular workshops and meetings and invite experts from the United States and abroad (e.g. Academic Sinica in Taiwan and Hong Kong Science and Technology University) to discuss open questions, share resources and seek consensus. We also visited China and Taiwan to present our work and ask for feedback.
- Annotators are encouraged to ask questions during the annotation process and in the second pass of bracketing randomly selected files are re-annotated by both annotators to evaluate their consistency and accuracy. Annotation errors and inter-annotation inconsistencies can reveal places in the guidelines that need revision.

In an ideal situation, guidelines would be available before annotation begins. However, real data from a corpus are far more complicated and subtle than examples discussed in the linguistics literature and many problems do not surface until sufficient data have been annotated. In this project, we divided each phase of the annotation and guideline development into three stages:

1. Our original guideline drafts are based on corpus analysis, review of the literature, and consultation with experts in treebanking and Chinese linguistics. As we begin the first pass of the annotation process, these guidelines evolve gradually through the resolution of annotation difficulties and annotator inconsistencies.
2. After the first pass, the guidelines are partially finalized and when possible the corpus is automatically converted to be consistent with the new guidelines before the second pass begins;
3. In the second pass our quality control method (Section 7) is designed to strengthen the guidelines by revealing annotation procedures that require elaboration for more consistent bracketing. Fortunately, our necessary elaborations at this stage have been very few.
4. After the second pass, the guidelines are finalized and the annotation is revised if necessary.

[5] *EAGLES* stands for the *Expert Advisory Group on Language Engineering Standards*. For more information, please check out its website at http://www.ilc.pi.cnr.it/EAGLES/home.html.

In this project, through careful design of the first version of the guidelines, no substantial changes have been made in the following versions and most revision of the annotation is done automatically by simple conversion tools.

In the next section we discuss highlights from the segmentation and part-of-speech tagging annotation process, followed by a section on the bracketing annotation process.

5. Segmentation and Part-of-Speech Tagging

The first phase of corpus annotation is word segmentation and part-of-speech tagging.

5.1. Issues in Segmentation Guideline Preparation

The difficulty in defining the notion of *word* is not unique to Chinese,[6] but the problem is certainly more severe for Chinese for a number of reasons. First, Chinese is not written with word delimiters so segmenting a sentence into ”words” is not a natural task even for a native speaker. Second, Chinese has little morphological marking to ease word identification. Third, there is little consensus in the community on difficult constructions which could affect word segmentation. The handling of resultative verb compounds, for instance, depends on the analysis of the construction, for which there is still no consensus in the linguistics community. For example, one view on how a verb-resultative compound is formed says that a simple sentence with the compound is actually bi-clausal and the compound is formed by movement, therefore, the compound should be treated as two words. Another view believes the compound is formed in the lexicon, and therefore should be one word. Fourth, many monosyllabic morphemes which used to be able to stand alone become bound in Modern Chinese. The influence of Ancient Chinese makes it difficult to draw the line between bound morphemes and free morphemes, notions which could otherwise have been very useful for deciding word boundaries.

To test how well native speakers agree on word segmentation of written texts, we randomly chose 100 sentences (5060 *hanzi*) from the Xinhua newswire and asked the participants of the first CLP workshop to segment them according to their personal preferences.[7] We got replies from seven groups, almost all of whom hand corrected their output before sending it. Table 2 shows the results of comparing the output between each group pair. Here, we use three measures that are widely used to measure parsing accuracy: precision, recall, and the number of crossing brackets (Black et al., 1991).[8]

The experiment is similar to the one discussed in (Sproat et al., 1996) in which six native speakers were asked to mark all the places they might pause if they were reading the text aloud. In both experiments, the native speakers (or judges) were not given any specific segmentation guidelines. Following (Sproat et al., 1996), we calculate the arithmetic mean of the precision and the recall as one measure of agreement between each output pair, and the average agreement is 87.6%, much higher than 76% in (Sproat et al., 1996). Without comparing the data in these two experiments, we do not know for sure why the numbers differ so much. One factor that might have contributed to the difference is that the instructions given to the judges were not exactly the same: in our experiment, the judges were asked to segment the sentences into *words* according to their own definitions of a *word*, while in their experiment, the judges were asked to mark all places they might possibly pause if they were reading the text aloud. There are places in Chinese, e.g. between a verb and an aspect marker that follows the verb, where normally native speakers do not pause but they still treat the verb and the aspect marker as two words. Another factor that might explain why the degree of the agreement in our experiment was much higher is that in our experiment all the judges were well-trained computational linguists. Some judges had their own segmentation guidelines and/or segmenters. They either followed their guidelines or used their segmenters to automatically segment the data and then hand corrected the output. In either way, their outputs should be more consistent.

The fact that the average agreement in our experiment is 87.6% and the highest agreement among all the pairs is 91.5% confirms the belief that it is common for native speakers to disagree on where word boundaries should be. On the other hand, the average number of crossing brackets is only 5.4 and the lowest is 1. Furthermore, most of these crossing brackets later turned out to be caused by hu-

[6] Even for languages which use delimiters between words, such as English, the distinction between a *word* and a non-word is not always clear-cut. For example, *pro-* normally can not stand alone, therefore, it is like a prefix. However, it can appear in a coordinated structure, such as *pro- and anti-abortion*, and under the assumption that only words and phrases can be coordinated, it is like a word. For more discussions of different notions of words (e.g. morphological object, syntactic atom, phonological word and listeme), please refer to (Sciullo and Williams, 1987).

[7] We did not give them any segmentation guidelines. Though participants received no segmentation guidelines, some applied their own guideline standards for which they had automatic segmenters while others simply used their intuitions.

[8] Given a candidate file and a Gold Standard file, the three metrics are defined as: the precision is the number of correct constituents in the candidate file divided by the number of constituents in the candidate file, the recall is the number of correct constituents in the candidate file divided by the number of constituents in the Gold Standard file, and the number of crossing brackets is the number of constituents in the candidate file that cross a constituent in a Gold Standard file.

If we treat each word as a constituent, a segmented sentence is similar to a bracketed sentence and its depth is one. To compare two outputs, we chose one as the Gold Standard, and evaluated the other output against it. As noted in (Sproat et al., 1996), for two outputs J_1 and J_2, taking J_1 as the Gold Standard and computing the precision and recall for J_2 yields the same results as taking J_2 as the Gold Standard and computing the recall and the precision respectively for J_1. However, the number of crossing brackets when J_1 is the standard is not the same as when J_2 is the standard. For example, if the string is *ABCD* and J_1 segments it into *AB CD* and J_2 marks it as *A BC D*, then the number of crossing brackets is 1 if J_1 is the standard and the number is 2 if J_2 is the standard.

	1	2	3	4	5	6	7	average
1	-	90/88/6	90/90/4	83/88/3	92/91/3	91/91/3	92/84/9	90/89/5
2	88/90/3	-	87/90/3	80/88/14	89/90/4	86/89/3	89/83/7	87/88/6
3	90/90/3	90/87/5	-	82/88/2	89/88/5	89/89/4	89/82/10	88/87/5
4	88/83/9	88/80/10	88/82/7	-	92/86/7	86/81/9	87/74/16	88/81/10
5	91/92/3	90/89/4	88/89/4	86/92/9	-	90/90/4	92/85/8	90/90/5
6	91/91/3	89/86/6	89/89/4	81/86/3	90/90/4	-	91/83/10	89/88/5
7	84/92/1	83/89/2	82/89/2	74/87/4	85/92/1	83/91/1	-	82/90/2

Table 2: Comparison of word segmentation results from seven groups

man errors. This implies that much of the disagreement is not critical and if native speakers are given good segmentation guidelines, the agreement between them will improve greatly.

So what is a word? The following tests for establishing word boundaries have been proposed by various authors: (Without loss of generalization, we assume the string that we are trying to segment is X-Y, where X and Y are two morphemes)

- bound morpheme: a bound morpheme should be attached to its neighboring morpheme to form a word when possible.

- productivity: if a rule that combines the expression X-Y does not apply generally, i.e., it is not productive, then X-Y is likely to be a word.

- frequency of co-occurrence: if the expression X-Y occurs very often, it is likely to be a word.

- complex internal structure: strings with complex internal structures should be segmented when possible.

- compositionality: if the meaning of X-Y is not compositional, it is likely to be a word.

- insertion: if another morpheme can be inserted between X and Y, then X-Y is unlikely to be a word.

- XP-substitution: if a morpheme can not be replaced by a phrase of the same type, then it is likely to be part of a word.

- the number of syllables: several guidelines (Liu et al., 1993; Chinese Knowledge Information Processing Group, 1996) have used syllable numbers on certain cases. For example, in (Liu et al., 1993), a verb-resultative compound is treated as one word if the resultative part is monosyllabic, and it is treated as two words if the resultative part has more than one syllable.

All of these tests are very useful. However, none of them is sufficient by itself for covering the entire range of difficult cases. Either the test is applicable only to limited cases (e.g. the XP-substitution test) or there is no objective way to perform the test as the test refers to vaguely defined properties (e.g. in the productive test, it is not clear where to draw the line between a *productive* rule and a *non-productive* rule). For more discussion on this topic from the linguistics point of view, please refer to (Packard, 1998; Sciullo and Williams, 1987).

Since no single test is sufficient, we chose a set of tests for our segmentation guidelines which includes all of the ones mentioned except for the productivity test and the frequency test. Rather than have the annotators try to memorize the entire set and make each decision from these principles, in the guidelines we spell out what the results of applying the tests would be for all of the relevant phenomena. For example, for the treatment of verb-resultative compounds, we select the relevant tests, in this case the number of syllables, the insertion test, and the XP-substitution test, and give several examples of the results of applying these tests to verb-resultative compounds. This makes it straightforward, and thus efficient, for the annotators to follow the guidelines.

The guidelines are organized according to the internal structure of the corresponding expressions (e.g. a verb-resultative compound is represented as V+V, while a verb-object expression is as V+N), so it is easy for the annotators to search the guidelines for needed references. The segmentation guidelines, including the comparisons between our guidelines and the ones used in China and Taiwan, can be found on our website.

5.2. Issues in POS Tagging Guidelines

Since Chinese words are not marked with respect to tense, case, and number, the central issue in POS tagging is whether the definitions of POS tags should be based on meaning or on syntactic distribution. This issue has been debated since the 1950s (Gong, 1997) and there are still two totally different viewpoints. For example, a word such as 毁灭 in Chinese can be translated into *destroy/destroys/destroyed/destroying/destruction* in English and it is used the same way as its translations in English. According to the first view, POS tags should be based solely on meaning. Since the meaning of the word remains roughly the same across all of these usages, it should always be tagged as a verb. The second view says POS tags should be determined by the syntactic distribution of the word. When 毁灭 is the head of a noun phrase, it should be tagged as a noun in that context. Similarly, when it is the head of a verb phrase, it should be tagged as a verb.

We have chosen the second view since it complies with the principles adopted in contemporary linguistics theories, such as the notion of head projections in X-bar theory and GB theory, and since it emphasizes the similarity between Chinese and other languages.

One argument that is often used against the second view is that since many verbs in Chinese can also occur in noun positions, thus requiring two POS tags, this increases the size of the lexicon. However, the extra POS tag allows us to distinguish between these verbs and many other verbs (such as monosyllabic verbs) which can not occur in noun positions. In addition, if there are generalizations about which verbs can occur in noun positions and which can not, these can be represented as morphological rules which allow the lexicon to be expanded automatically. On the other hand, if no such generalizations exist and the nominalization process is largely idiosyncratic, it supports the view that this is a lexical phenomenon and verbs which can be nominalized should be marked by having two POS tags in the lexicon. Finally, the phenomenon that many verbs can occur in noun positions is not unique to Chinese, and the standard treatment in other languages is to give them both tags.

5.3. Annotation Process

Before the first pass of annotation, we finished a draft of segmentation and POS tagging guidelines. Our corpus was automatically segmented and POS tagged by the BBN/GTE integrated stochastic segmenter and part-of-speech tagger. The tagger was trained on the Academia Sinica Balanced Corpus (ASBC). Since the ASBC guidelines and our guidelines have some differences (cf. our website), we wrote tools to convert the ASBC tags into our tags automatically. Although the mapping was not one-to-one and introduced some errors, this process greatly accelerated annotation.

The first pass (including the time spent on training annotators) took roughly four months to complete. After the first pass, the guidelines were revised and the second pass began, where one annotator double-checked the files annotated by the other annotator. The second pass took less than two months because the annotators were well-trained by then and the input of the second pass was much better than the input of the first pass. To identify possible tagging errors, we compiled a list of sorted (word, POS tag) pairs and checked the list for implausible tags. The POS tagged corpus has 100,991 word tokens and 10829 word types. The number of unique (word, POS tag) pairs is 11960. Therefore the average number of POS tags per word is only 1.10 (11960 divided by 10829). Counting only the words that occur more than once, the number increases from 1.10 to 1.21.

Our tagset with 33 POS tags enabled us to capture syntactic structure with minimal tagging complexity. Analysis reveals that our corpus approaches lexical-semantic closure at a rate which can be favorably compared with the Chinese newspaper sub-corpus in the ASBC. By the time 99,000 tokens of our corpus are tagged, only 23 new token+tag combinations were observed in the last 1,000 tokens and half (12) of those are new proper nouns. In comparison, 57 new token+tag combinations were observed in the last thousand of 99,000 tokens in the ASBC ”A*” newspaper sub-corpus and 37 of those were new nouns.

6. Syntactic Bracketing

The second phase of corpus annotation is syntactic bracketing.

6.1. Issues in Guideline Preparation

This section discusses three issues we addressed when creating our bracketing guidelines. The first issue is the choice of a representation scheme. Given that the sentences in the corpus are very long and complex, the representation scheme needs to be robust enough to be able to represent all the important grammatical relations and at the same time be sufficiently simple so that the annotators can follow it easily and consistently. An overly complicated scheme will slow down productivity and jeopardize consistency and accuracy. In our representation scheme, each bracket has a syntactic label and zero or more functional tags. The label indicates the syntactic category of the phrase, while the function tags provide additional information. For example, when a noun phrase such as 昨天*/yesterday* modifies a verb phrase, its syntactic label will be *NP* (for noun phrase) and it is given a function tag *-TMP*, indicating that the NP is a temporal phrase and its function is similar to that of an adverbial phrase. We also use reference indices to mark syntactic movement. Our scheme is similar to the one adopted in the Penn English Treebank (Marcus et al., 1994).

The second issue is the treatment of various syntactic constructions. Many of them, such as the *ba*-construction and the *bei*-construction, have been investigated for decades, but there is still no consensus on how they should be analyzed. To tackle this issue, we: (1) studied the linguistics literature, (2) attended Chinese linguistics conferences, (3) had discussions with our linguistic colleagues, (4) studied and tested our analyses on the relevant sentences in our corpus, and (5) used special tags to mark crucial elements in these constructions. For example, the word 把 in the *ba*-construction has been argued to be a case marker, a secondary topic marker, a preposition, a verb, and so on in the literature. Clearly, the word is different from other prepositions and other verbs and there is no strong evidence to support Chinese having overt case markers or topic markers. We believe the word is more like a verb than a preposition, but to distinguish it from other verbs, we assign it a unique POS tag *BA* and in the bracketing guidelines we give detailed instructions on how to annotate the construction. If some users of our corpus prefer to treat it as a preposition, it is easy to convert our annotation to accommodate that approach.

The third issue with respect to bracketing guidelines is the treatment of ambiguous sentences. In the guidelines, we have classified ambiguous sentences according to the origins of their ambiguity, and specified the treatment for each type. For example, a subset of Chinese adverbs can occur either before the subject or after it. When the subject is phonologically empty as a result of pro-drop or relativization, the empty subject can be marked either before the adverb or after it without much difference in meaning and there is no syntactic evidence to favor one analysis over another. If nothing is specified in the guidelines and the annotator is allowed to mark the empty subject in either place, there will be inconsistency, even though both analyses are plausible. In this case we specify a ”default” position for the subject and require that the empty subject be put before the adverb. By doing so we avoid the need for annotators to make individual choices, which is a potential cause of inconsistency.

6.2. Annotation Process

The first pass of bracketing started once we had finished a preliminary draft of the bracketing guidelines and it took two annotators four months. The primary goal of this pass was to identify the problems in the draft guidelines such as certain constructions that were not covered or analyses that could not be extended to account for new data. Having a corpus bracketed, albeit coarsely, also made it possible for us to use a corpus search tool that operates on bracketed sentences. With this tool we were able to pull out all the sentences in the corpus relevant to a particular construction and come up with better generalizations that are more robust in accommodating new data.

After the first pass, we did an extensive revision of the bracketing guidelines based on the problems we had collected and solved during the first pass. After training the annotators on the revisions the second pass began. The emphasis of the second pass was quality control, i.e. to ensure inter-annotator consistency and annotators’ adherence to the guidelines as discussed in the next section. After the second pass, we plan to run the corpus search tool and a grammar extraction tool (Xia, 1999). Both tools are able to identify certain types of annotation errors which will facilitate our final cleanup of the corpus and final revisions of the three sets of guidelines. Following the completion of the bracketing annotation and guidelines, we will release the corpus and guidelines to the public and organize a second Chinese Language Processing (CLP) workshop in conjunction with ACL’00.

7. Quality Control

A major challenge in providing syntactic annotation of corpora is ensuring consistency and accuracy. The approach that we take is a pragmatic one and is derived from the constraints imposed in building a Chinese Treebank without any access to a reference grammar (Black et al., 1993), or any existing parsing support tools (Marcus et al., 1993; Black et al., 1996; Chen and Shaw, 1998; Kurohashi and Nagao, 1998; Skut et al., 1998). However, the consistency we have attained as described below validates our manual method with automatic evaluation.

Carefully documented guidelines, linguistically trained annotators, and annotator support tools are pre-requisites to creating a high quality corpus with acceptable production rates. Section 4 above describes our methodology for creating the guidelines. Both of our annotators are linguistics graduate students, one of whom authored the Bracketing Guidelines and regularly participated in the meetings on Chinese syntax. Their knowledge of linguistics, in general, and syntax, in particular, is crucial for the success of the project. To support bracketing, our annotators use the bracketing interface described in Marcus et al.(1993) and a corpus search tool for linguistic investigation and comparison of annotations.

With the first pass complete and the annotation guidelines researched and documented as much as possible, we adopted a quality control method. Our goal was to accelerate the annotation of consistent and accurate data by eliminating the need for blind double re-annotation[9] of the entire corpus, yet account for annotator consistency and adherence to guidelines throughout the second pass. A secondary goal is to create a subset of data for testing (20% of the corpus) that had double re-annotation and could serve as a Gold Standard.

Our primary tool for evaluating consistency is the Parseval software that produces three metrics — precision, recall and numbers of crossing brackets (Black et al., 1991), which we extended to process Chinese. Our process of evaluating consistency is as follows: first, some files from the output of the first pass are randomly selected for double re-annotation. Next, Parseval is used to compare the two independent re-annotations, and any discrepancies are carefully examined and the annotation is revised. This may in turn lead to revisions of the guidelines to prevent a recurrence of similar inconsistencies. Then, the corrected, reconciled annotation is considered the Gold Standard, and each of the two original re-annotations is then run against it and against each other, again using Parseval, to provide a measure of individual annotator accuracy and inter-annotator consistency. Although the Parseval scheme may not be ideal for evaluating a diverse range of parsers at variance with the reference corpus (Gaizauskas et al., 1998), for our purposes the scheme quantifies and evaluates the two annotators’ parses and allows us to track accuracy and consistency.

We first used this method to re-train annotators at the beginning of the second pass, when forty files from the output of the first pass were randomly selected for double re-annotation. After that, the annotators continued to correct first pass data and each week two files were randomly selected and double re-annotated and the re-annotations were compared with Parseval software. In this way, we continue to monitor our consistency and accuracy and to enhance guidelines. Table 3 shows the accuracy of each annotator (denoted by *1st* and *2nd* in the table) compared to the Gold Standard and the inter-annotator consistency in the first four weeks after the re-training period. The table shows both measures are in the high 90% range, which is more than satisfactory.

In addition to the Parseval software, we also use a corpus search tool and a grammar extraction tool to extract patterns which can be inspected to pinpoint certain types of

[9] Double re-annotation means both annotators re-annotate the same files from the output of the first pass.

	Accuracy				*Consistency*	
	1st vs. gold		2nd vs. gold		1st vs. 2nd	
Week	prec	recall	prec	recall	prec	recall
1	98.15	97.58	95.56	96.41	94.26	95.67
2	97.85	98.95	97.71	98.41	95.82	97.60
3	96.21	97.56	95.18	96.58	92.05	94.90
4	95.48	97.86	96.08	93.97	92.39	92.62
Avg	96.92	97.99	96.13	96.34	93.63	95.20

Table 3: Accuracy and inter-annotator consistency for the first four weeks after the re-training period

annotation errors.

Our method is a new position along the scale of human-to-automatic processing described in (Bateman et al., 1997). We begin with human analysis and input, conduct human analysis enhanced by an available editor, and apply automatic evaluation and human analysis with correction and completion by human post-editors.

8. Conclusion

We have discussed in detail the approach we have used to create a 100K word Chinese Treebank, including the development of the guidelines for segmentation, POS tagging and syntactic bracketing, as well as our methodology for ensuring inter-annotator consistency and community involvement. We look forward to distributing our annotated corpus and getting feedback from the community on its usefulness. We think our methodology for guideline development and consistency checking will be applicable to monolingual text annotation for other languages as well, and will be testing this hypothesis.

9. Acknowledgement

We gratefully acknowledge the invaluable assistance of BBN, ROCLING, and Hong Kong Science and Technology University in the guideline preparation and annotation. We thank Zhibiao Wu for porting our treebanking tools to Chinese and assisting with encoding issues and Ann Taylor for her extremely pertinent treebanking advice. We also thank Andi Wu and Zixin Jiang from Microsoft, Jin Yang from Systran, Shiwen Yu and Hui Wang from Peking University, Changning Huang from Tsinghua University, and our many linguistics consultants including Peter Cole, Shengli Feng, James Huang and Aravind Joshi. Finally we thank all of the participants at our meetings and workshops.

10. References

Bateman, J., J. Forrest, and T. Willis, 1997. The use of syntactic annotaion tools: partial and full parsing. In R. Garside, G. Leech, and A. McEnery (eds.), *Corpus Annotation*. London: Longman.

Black, E., S. Abney, D. Flickinger, C. Gdaniec, and et. al., 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop*.

Black, E., S. Eubank, H. Kashioka, D. Magerman, R. Garside, and G. Leech, 1996. Beyond skeletal parsing: producing a comprehensive large-scale general English treebank with full grammatical analysis. In *Proceedings of COLING*. Copenhagen, Denmark.

Black, E., R. Garside, and G. Leech, 1993. *Statistically-Driven Grammars of English IBM/Lancaster Approach*. Rodopi Editions:Amsterdam.

Chen, H-H and M-S Shaw, 1998. A treebank development tool. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada (eds.), *First International Conference on Language Resource and Evaluation*.

Chinese Knowledge Information Processing Group, 1996. Shouwen Jiezi - A study of Chinese Word Boundaries and Segmentation Standard for Information Processing (In Chinese). Technical report, Taipei: Academia Sinica.

Feng, Shengli, 1998. Short Passives in Modern and Classical Chinese. In *The 1998 Yearbook of the Linguistic Association of Finland (41-68)*.

Gaizauskas, R., M. Hepple, and C. Huyck, 1998. A scheme for comparative evaluation of diverse parsing systems. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada (eds.), *First International Conference on Language Resource and Evaluation*.

Gong, Qianyan, 1997. *zhongguo yufaxue shi (The history of Chinese syntax)*. *Yuwen* Press.

Kurohashi, S. and M. Nagao, 1998. Building a Japanese parse corpus while improving the parsing system. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada (eds.), *First International Conference on Language Resource and Evaluation*.

Liu, Y., Q. Tan, and X. Shen, 1993. Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology.

Marcus, M., B. Santorini, and M. A. Marcinkiewicz, 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Lingustics*.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, et al., 1994. The Penn Treebank: annotating predicate argument structure. In *In Proc of ARPA speech and Natural language workshop*.

Packard, Jerome L. (ed.), 1998. *New Approaches to Chinese Word Formation*. Mouton de Gruyter.

Sciullo, Anna Maria Di and Edwin Williams, 1987. *On the definition of word*. The MIT Press.

Skut, W., T. Brants, B. Krenn, and H. Uskarect, 1998. A linguistically intepreted corpus of German newspaper text. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada (eds.), *First International Conference on Language Resource and Evaluation*.

Sproat, R., W. Gale, C. Shih, and N. Chang, 1996. A Stochastic Finite-state Word Segmentation Algorithm for Chinese. *Computational Linguistics*.

Xia, Fei, 1999. Extracting tree adjoining grammars from bracketed corpora. In *Proceedings of 5th Natural Language Processing Pacific Rim Symposium (NLPRS-99)*. Beijing, China.